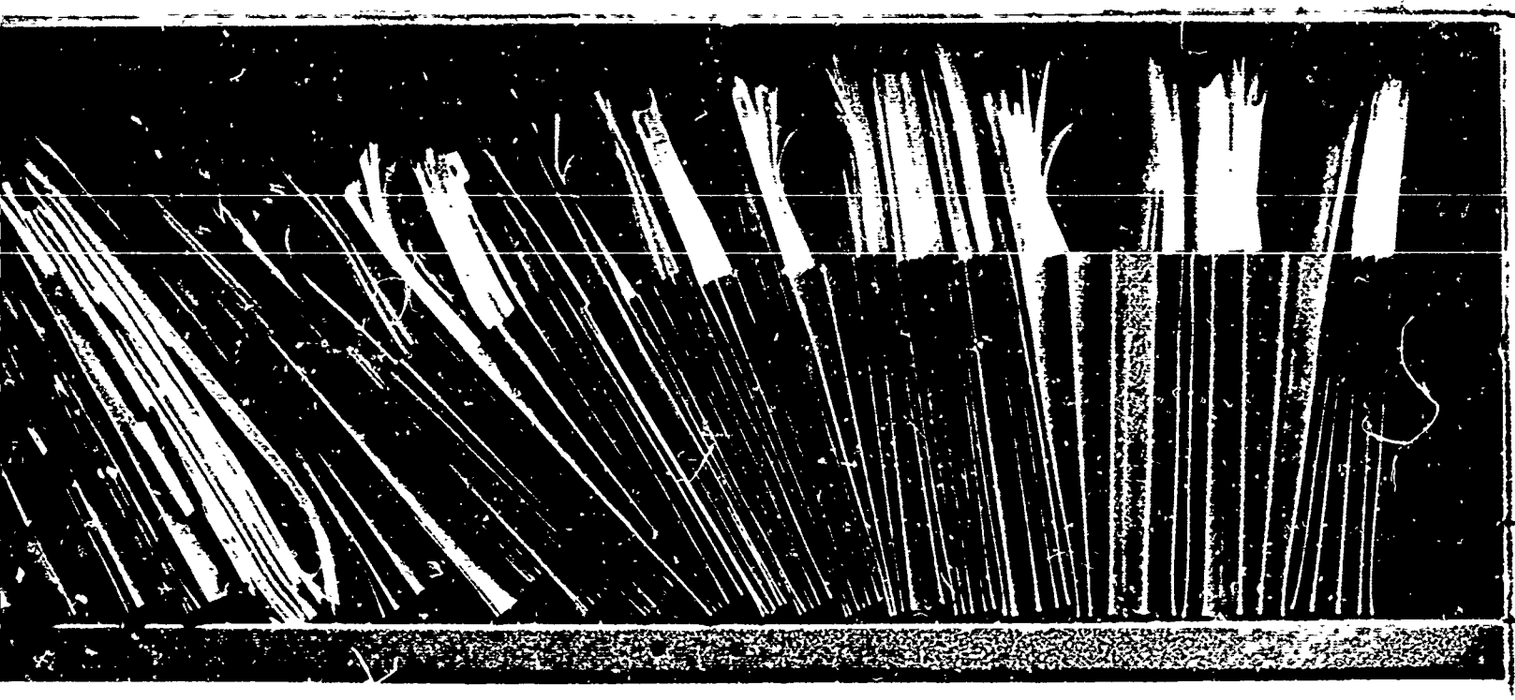
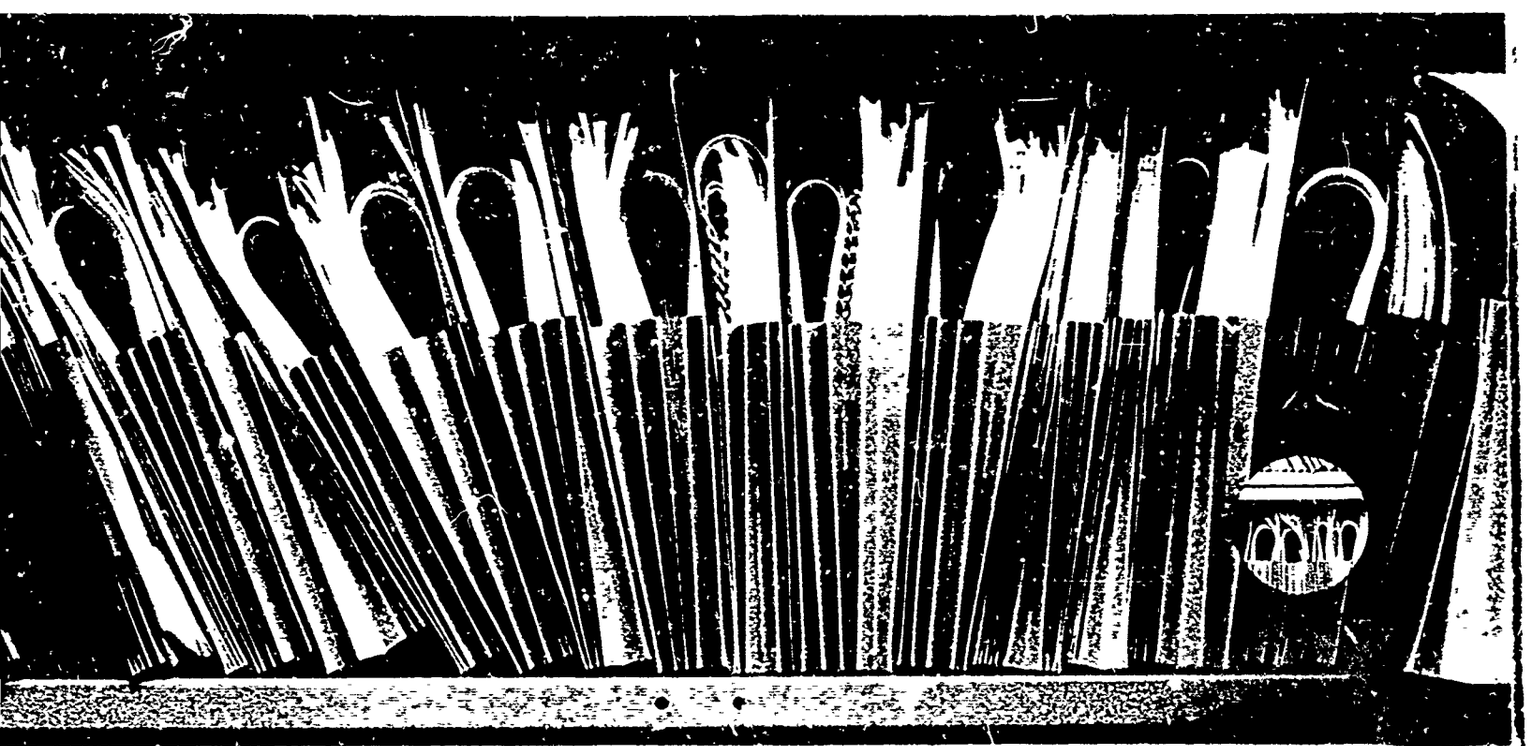
System Development Corporation

Technology Series

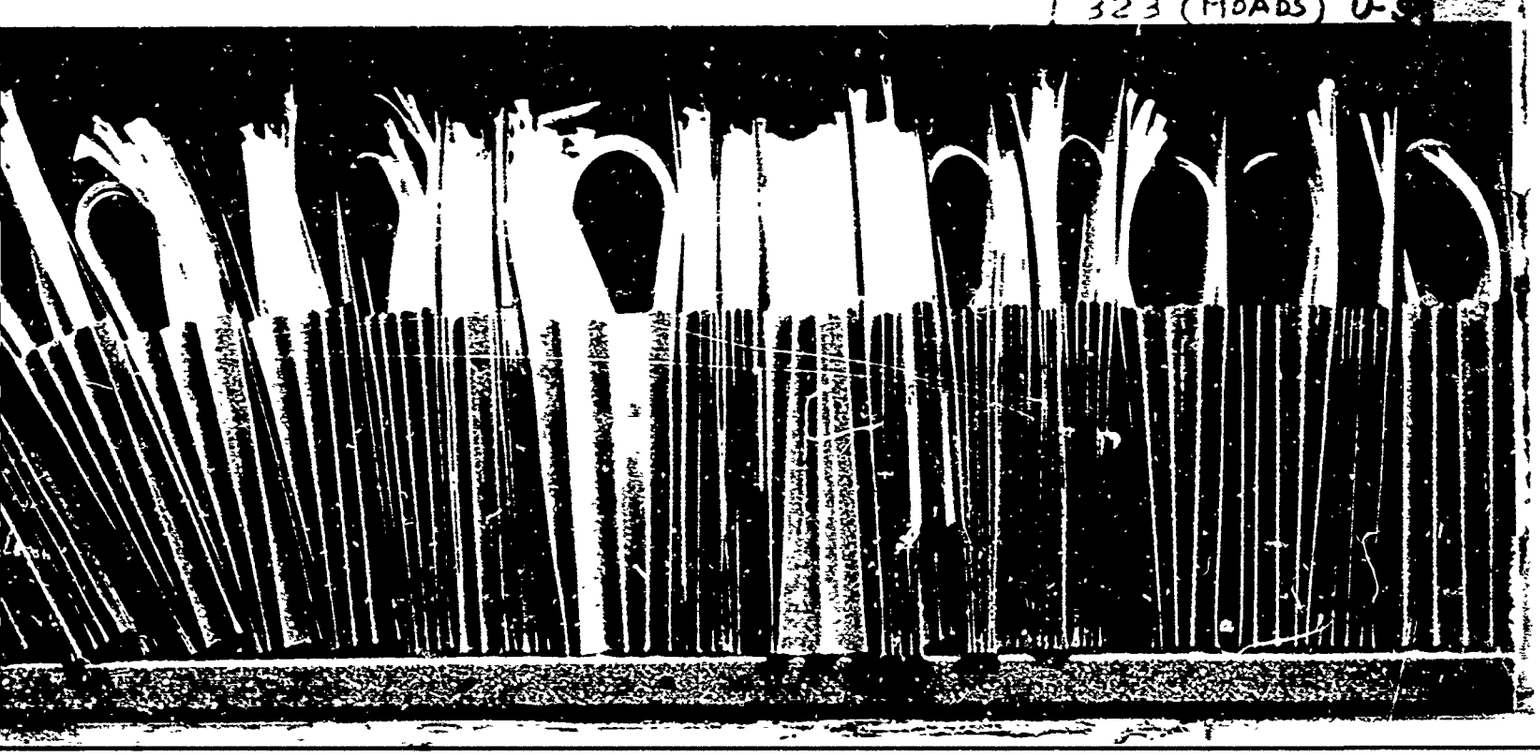
Information Retrieval

INFORMATION RETRIEVAL





1323 (MOADS) U-S



THE INFORMATION PROBLEM

It has become a truism in recent years to say that science and technology, government, business and industry, in fact, all the professional fields, are confronted with an acute problem in information. the vast amounts of data being generated in what has been described as an "information explosion" that are in critical need of organization for rapid and accurate accessibility.

In addition to a multiplicity of private organizations working in the field, the United States government has shown increasing concern. Committees of both houses of Congress have explored the problem. A special panel of the President's Science Advisory Committee spent more than a year studying the responsibilities of the technical community and the government and identifying the problems in information handling that have been magnified by the accelerating growth of science and technology

PARTICIPATION IN THE FIELD

Millions of dollars a year are already being spent by industry as well as government agencies on the development of assorted information retrieval systems. System Development Corporation has been directly concerned as a developer and user of information and as an organization that assists

government and educational organizations in the effective use of advanced information processing techniques. Its work in the specific field of information retrieval dates back to 1958.

AN AGE-OLD PROBLEM

Despite the current emphasis on the need for information retrieval systems, the problem of storing and retrieving knowledge is an ancient one. It could probably be traced back to the cave paintings made some 25,000 years ago when, unknowingly, man took the first step toward recording experience and communicating it beyond the range of his own voice and the span of his own lifetime.

EVOLUTION OF INFORMATION RETRIEVAL

For thousands of years thereafter libraries acquired, organized, stored, and made accessible on demand the written products of human experience; information retrieval, as a modern concept, was first thought of in terms of simply storing and retrieving documents. However, the growing volume of publications—in 1960 the world's technical literature was estimated at one to two million papers a year appearing in 30,000 journals—has pressured the technical community to devise ingenious schemes for document retrieval. But document retrieval alone is not enough: a technical specialist really needs specific information contained within documents, not merely the undigested bulk of a body of published literature.

As a result, the problem of information retrieval has become considerably more sophisticated. It involves not only the development of advanced equipment, but also profound conceptual questions on the definitions of "information," "fact," and the various uses of both.

HARDWARE

Great advances have been made in solving many of the physical aspects of the problem. A variety of hardware devices is readily available, ranging from simple perforated card equipment to elaborate digital computers and

special-purpose search and processing gear. Vast improvements have been made in storage mechanisms and there has been significant progress in the development of devices, such as optical character readers, that will permit the rapid and economical preparation of information for machine processing.

TECHNIQUES

The most difficult problems are posed by basic concepts. A major challenge —though it appears deceptively simple—is learning how to label information items so that they are not only adequately described, but also differentiated from other, possibly similar items. Another critical problem is learning how to help the user to ask the right questions. Often he is unable to state precisely the information he is seeking and his initial search request does not prove fully productive. Systems are needed that can describe their data holdings to the user in such a way that he is able to browse for leads and search paths which may help him locate specific items of interest.

There are also serious difficulties in the area of information distribution, or dissemination. Specifically, the problem is to characterize the interests of individual information users in such a way that pertinent information can be selected from a stream of documents and brought immediately to their attention.

INFORMATION MANAGEMENT SYSTEMS

In the process of broadening the basic concepts of the field, the term "information retrieval" has become too restrictive to encompass the full scope of system functions. The more inclusive designation of "information management" has been adopted at SDC.

INFORMATION MANAGEMENT

What is information management? Stated most simply, it is the establishment and utilization of effective procedures for controlling the generation, processing, flow and use of information. Procedures can be manual or machine-aided, although they are most commonly considered as involving machine aids.

20
21
22
23
24
25
27
28
29
30
31
32
33
34
35

HISTORICAL DEVELOPMENT

DOCUMENTATION

In the SDC view, information management technology is the outgrowth of three separate streams of activity. The first is that of documentation, best represented by the library. The object of interest of this field has traditionally been the document—a book or treatise—and the emphasis has been largely on archival operations. Timeliness has not been an overriding consideration. Much of the literature on the subject of information retrieval and automated literature searching has been generated by the documentation activity.

BUSINESS SYSTEMS

A second major contribution to information management technology has come from the area of business operations. Since the late 1930's, machines have been used extensively to expedite tasks connected with sorting, storing, and retrieving recorded business information. Insurance records management is typical of this type of application. The business community has been primarily responsible for pioneering the use of large semiautomated data files. The information in these files, unlike that with which libraries are concerned, is principally numerical, generally concise and usually requires fairly timely processing to be of value. Some government

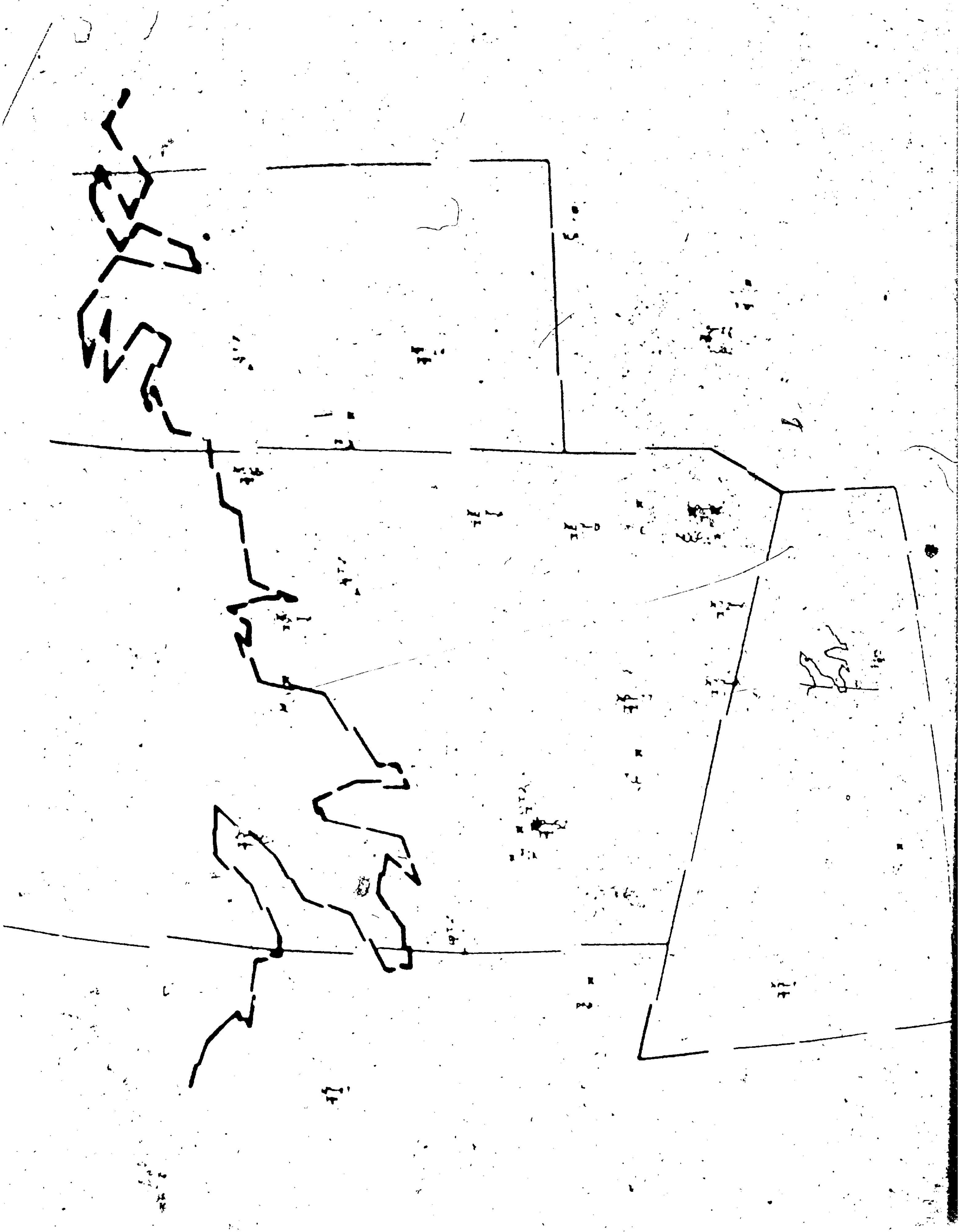
operations, like the maintenance of labor statistics and census data, may be considered part of the same stream of activity.

MILITARY SYSTEMS

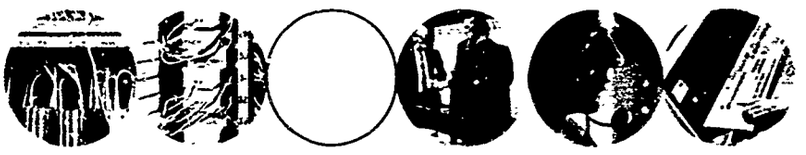
In the middle of the 1950's a third major influence became significant. The digital computer, which provided faster and more powerful information-handling capabilities than had ever been available before, began to be used extensively as the central element of large real-time military information processing systems. The first and largest of these was the SAGE (Semi-Automatic Ground Environment) air defense system, which used a network of large digital computers to process thousands of air defense events on virtually an instantaneous basis.

Computer-based military systems, like the business records systems, work primarily on concisely stated, well-formatted information. They differ from business systems primarily in terms of the complexity of the operations they perform and in their extreme speed of operation. For example, SAGE filters and organizes thousands of bits of data every few seconds, distributes it selectively to individual operator positions, collates and compares new and stored data, and displays different aspects of the air picture on command. The system also provides the human operators with considerable assistance in exploring decision situations and the choice of alternative courses of action.

Modern-day information management systems reflect these three streams of activity and utilize many elements of each. The enormous range of possible system requirements poses very difficult challenges for the equipment manufacturer, the developer of information-handling techniques, the designer of information management systems, and the user of these systems. Current systems must often be geared to deal both with natural language and with numerical data; with complete documents and with individual facts; with archival-type storage and with small dynamically changing files; with immediate information distribution and with retrospective search; with broad generic searches and with spot questions. It is clear that the current state-of-the-art has yet to provide completely satisfactory solutions to these problems.



MILITARY APPLICATIONS



SAC CONTROL SYSTEM

The information management aspects of the SAGE System—for which SDC has done the computer program design and development—have already been mentioned. Information management plays an important role in another major computer-based military system on which SDC is working, the Strategic Air Command Control System (SACCS). To operate effectively, SAC requires current information on the condition and disposition of its global forces, world-wide weather conditions, and the location of critical enemy forces and facilities.

The primary form of response to interrogation in the SAC Control System is through displays—that is, outputs whose format is predetermined, but whose content is variable. These displays, developed from basic data kept in the computer's peripheral storage equipment and continually updated by inputs from SAC units and other systems, are designed to meet the known information needs of SAC decision-makers.

In the SAC Control System, as in documentation systems, it is exceedingly difficult, if not impossible, to anticipate all information needs. Instances occur in which the basic information has been stored, but no specific output

for it has been designed. As part of its work on this system, SDC has developed data retrieval programs which allow the SAC user to request and receive printouts in various formats of any data defined in the system's data base.

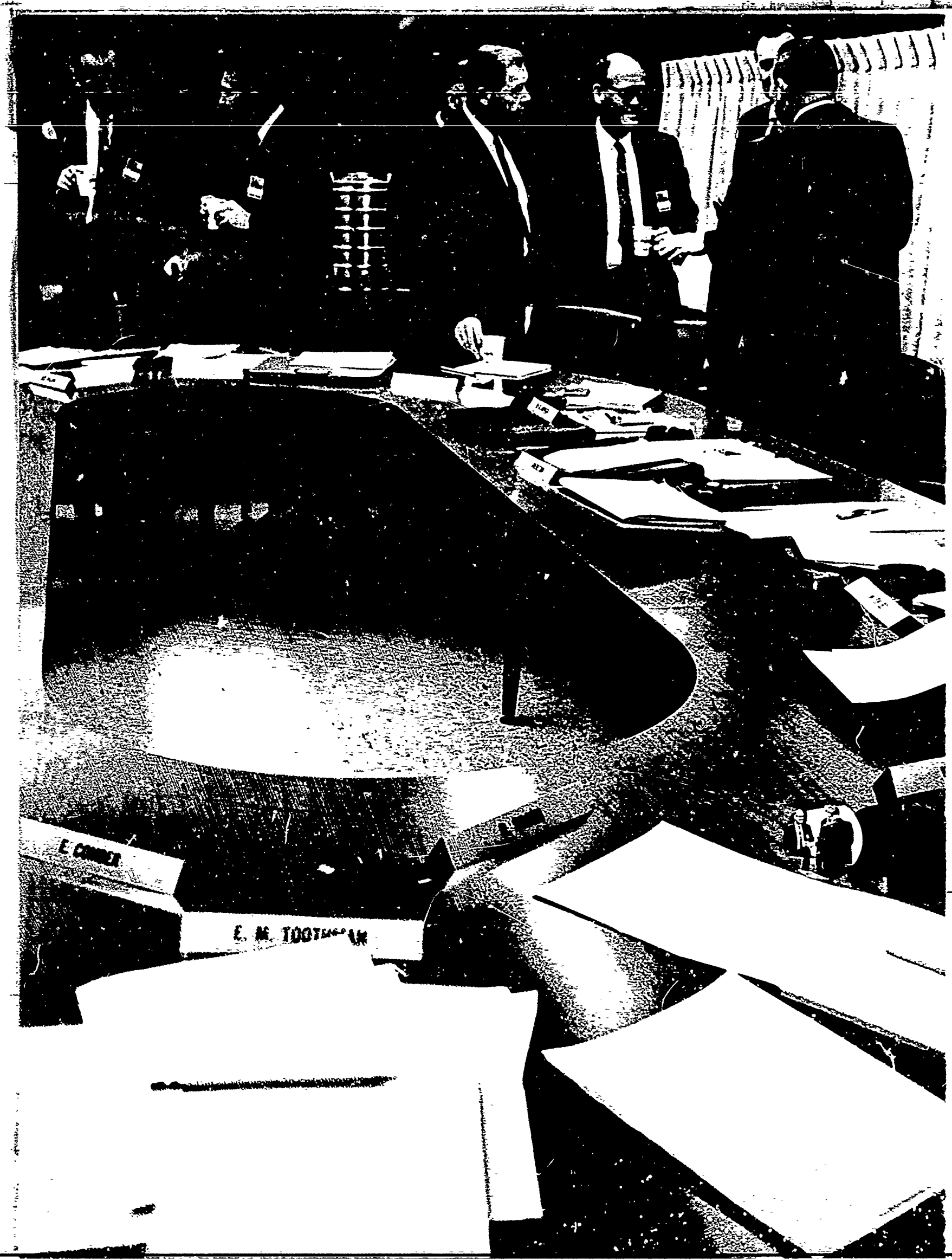
INTELLIGENCE SYSTEMS

Intelligence systems present particularly complex information management problems because these systems, more than the military command or control systems, deal primarily with natural language data. Many of these systems must cope with such diverse language materials as foreign journal and newspaper articles, translations, abstracts, brief teletype messages and standard intelligence reports. The range of formats, variety and content, and the requirements for data correlation and for access at many levels of generality and specificity have required the design of unique thesaurus-building programs and other language processing procedures.

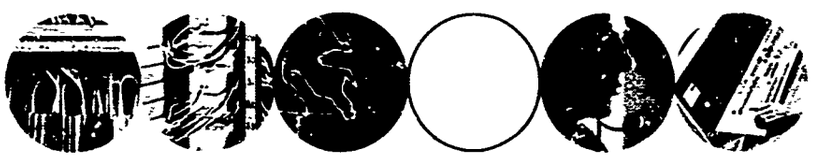
DATA BASE MANAGEMENT

In most large information management systems, organization of the data base—the data on which the system operates—is an extremely important activity. Data bases vary in storage media and in the characteristics of their items. The ease with which the data base can be "loaded" and the flexibility with which it can be searched, manipulated and updated determine to a large extent the operational capabilities of the system.

In its work for the Department of Defense Damage Assessment Center, SDC developed a data base system to store and process the extremely large repository of data needed for rapid damage assessment. The problem was complicated by the fact that much of the data, which had to be obtained from other systems, was in a variety of formats. It was necessary to store the information in the form required by the user's system, as well as organize it for efficient utilization. The data base system designed to meet this requirement is parameter-controlled. Thus it provides the flexibility necessary for constructing, loading and manipulating the information from these different sources. Further, the system allows the user to specify the required actions in understandable English.


E. CONNER
E. M. TOOTHMAN

NONMILITARY APPLICATIONS



Equally as interesting as the computer-based military systems are the newly emerging information systems in the fields of medicine, education, urban planning, law and scientific and technical documentation.

MEDICINE

For more than four years, SDC has been engaged in research and development of information management systems for the medical sciences. Since 1960 SDC and the Veterans Administration (VA) have been collaborating on the development of an integrated information management system for the Department of Medicine and Surgery. SDC is also designing an information processing network for a vast medical center under construction in San Juan, Puerto Rico.

In one phase of the VA program, SDC established and operated a prototype medical research support center to determine manual and computer methods for recording, storing, retrieving, and analyzing research data. In another part of the program concerned with the handling of medical records, a computer-based system that rapidly accepts, processes and retrieves patient data in a simulated hospital setting was successfully developed and demonstrated.

EDUCATION

More and more school districts and institutions of higher learning are installing electronic data processing systems for the management of educational data. This development, coupled with the already extensive use of electronic accounting machines in the registrar and business offices, seems to portend a major technological change in school system operations. It is increasingly evident, too, that information processing techniques will extend into the classroom facilitating innovations in instructional techniques.

The research and development at SDC in the field of education has underscored the emphasis on automatic information processing. In 1960 SDC designed and constructed a computer-based laboratory to study automation in school systems. In this laboratory research has been conducted in the use of teaching machines, rapid information retrieval systems, and programmed TV lessons. The goal has been to develop the physical facilities, instructional programs, and information management techniques necessary to permit each student to progress through a curriculum at a rate consistent with his abilities and interest.

SDC has been engaged in defining and analyzing several school systems in the state of California, and in the case of one particular school, has been involved in designing the information processing system. The California State Department of Education and SDC are cooperating on the development of a computer model of a secondary school set in the context of a district, regional and state data-link network. This model will be used for further systems analysis and design work. SDC has also been working with the United States Office of Education (USOE) to analyze the data collection procedures

connected with several major educational programs. The intent of this study is to revise the information systems operations of the USOE and to establish a central data bank for the management of educational information.

METROPOLITAN GOVERNMENT

Metropolitan governments have complex problems in information management. These problems concern three major functions: operations, planning, and research. In their day-to-day operations, the individual city departments such as Public Works, Building and Safety, Health, Police, and Fire, as well as the different administrative agencies such as the city clerk, controller, and administrative officer, must deal with a great variety of information about businesses, real and private property, people, and events. The problems of acquiring, maintaining, and using this information have multiplied as a result of the rapid growth of metropolitan areas.

Urban planning has become singularly complex, with a pressing need for information on land use, facility requirements, and socioeconomic data. Information for urban research has also become an important requirement. The relationships among demographic behavior, economic characteristics, and residential, industrial, and transportation facilities of a metropolitan region are not well understood. Yet these relationships represent the base upon which metropolitan planning and operations rest.

One method for solving the information problems of metropolitan operations, planning, and research is to develop urban information systems that are integrated information management systems for given metropolitan regions. At present, SDC is conducting a study of a generalized urban information system, examining the information needs of decision-makers and other concerned individuals, potential system designs, cost/feasibility tradeoffs, and recommended steps for the development of a specific system.

LAW ENFORCEMENT

The problems of law enforcement information processing are becoming particularly severe as population expands and crime and traffic incidents occur at an ever faster rate. For some time, SDC has been working with the

law enforcement community . determine how information management technology can be applied to law enforcement problems.

A committee of the California Peace Officers Association is exploring the feasibility of a state-wide law enforcement information system to provide the field officer with rapid access to files containing wants and warrants for specific persons, stolen vehicle data, criminal records, and other identification information. SDC is participating in the work of this committee.

The other activity is concerned with the retrieval and analysis of crime information for purposes of general patrol distribution and investigation of specific crimes. In the processing of crime reports, present techniques employ precoded summaries abstracted from the original reports. Some potentially useful information is lost in the process. SDC is experimenting with natural language computer programs that will permit the storage and retrieval of crime reports in their original, full-text form.

LAW

The legal profession has used highly developed research aids for years, and over the past decade has turned increasingly to automated data processing. A project was begun in the spring of 1961 at SDC to determine the potential and the problems in applying automated data processing to the operations of trial courts in general and to the Los Angeles Superior Court in particular. To date the study shows that simple data processing aids can assist in several areas: filing, indexing, and maintaining case files; preparing statistical and management reports; managing the court calendar; determining diagnostic patterns in the events that lead to marital complaints; and processing of data related to those in custody. This project is a joint undertaking of the court, the UCLA Law-Science Research Center, and SDC.

A second project was begun in 1962 to investigate the utility of machine-prepared indexes of appellate decisions (using the word, "indexes," in a broad sense). It is felt that such indexes, when published, will provide those who cannot afford to use a computer with at least some of the benefits of computerized legal search. Experimental computer routines which index, concord, abstract, edit, or list several forms of concordances have been prepared.

SCIENTIFIC AND TECHNICAL INFORMATION

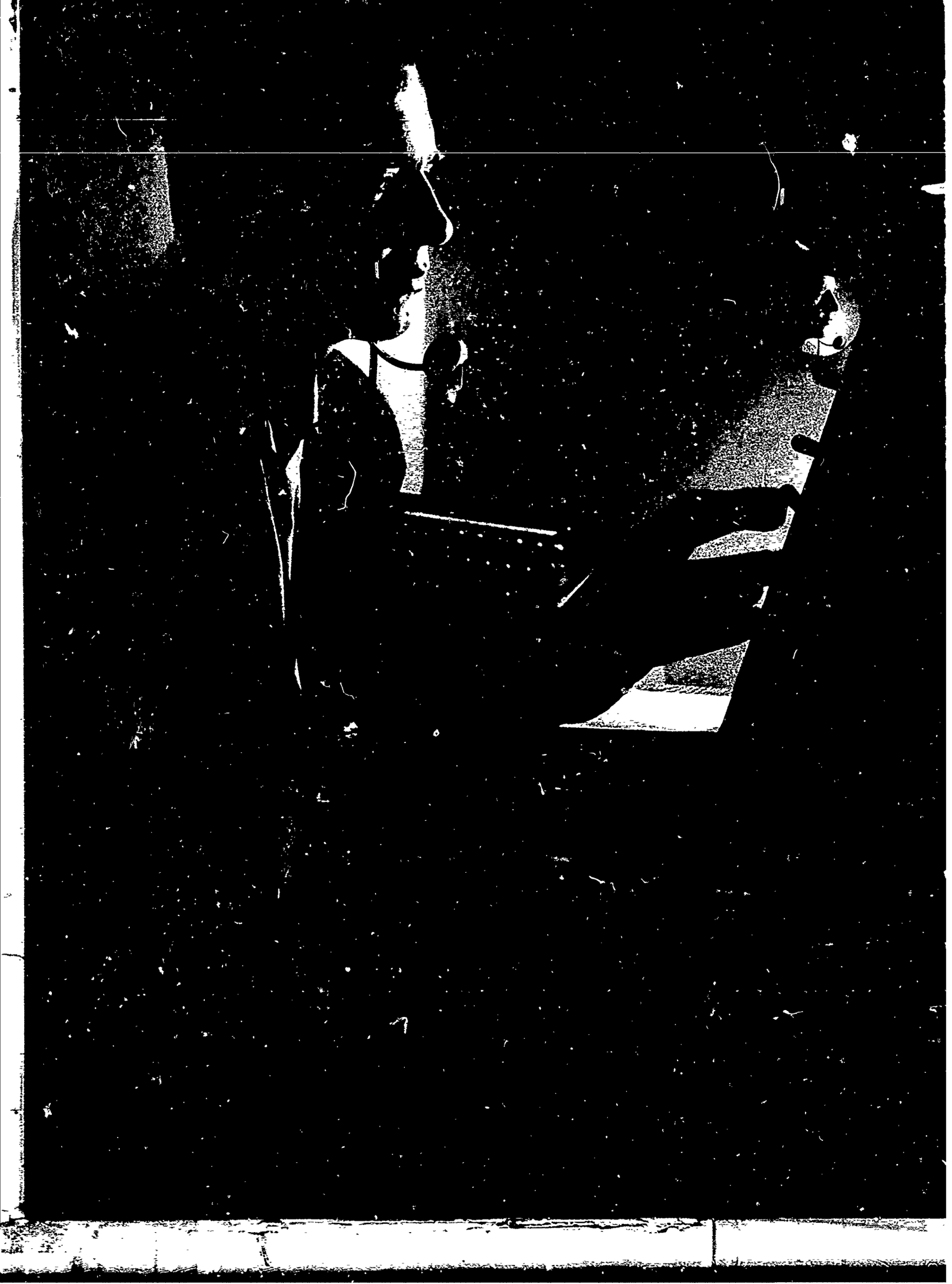
Probably the most urgently needed work in information management is in the area of scientific and technical information. The President's Science Advisory Committee concerned itself chiefly with this area and with the need to establish better management of the generation and flow of scientific and technical reports. SDC is working with the Science Information Exchange in Washington, D.C., exploring the use of the natural language of the reporting scientist to relate and retrieve research reports.

The Science Information Exchange maintains a clearing house for current research in the biological and physical sciences. Research reports from the current files of the exchange are being used in an experiment to compare full-text machine processing with present manual procedures for locating (i.e., retrieving) research reports related to specific queries. Machine-produced responses will be compared with actual operating results provided by the staff of the exchange. Computer programs developed at SDC for the handling of natural language data will be used extensively in this experiment.

IN-HOUSE DEVELOPMENT AND USE

Information-handling techniques are being applied in SDC's own operations. This use not only provides a valuable service, but also permits the techniques to be carefully evaluated in an operational setting.

An example of in-house development and use is the SATIRE system of Semi-Automatic Technical Information REtrieval. Developed several years ago, SATIRE is based on existing machine capabilities and techniques for document retrieval. Utilizing EAM equipment, SATIRE produces all library cross-reference card catalogs, preprinted cards for an automatic charge-out system, selected bibliographies, library indexes for publication, accession lists, and author inventories. A computer-based version of SATIRE yields the same products, but provides the advantage of tape storage, simplified updating and obsoleting procedures, a higher degree of accuracy, more complete indexes, and much greater speed of operation. The system, developed to serve both scientists and librarians, is in operation in two of SDC's divisions.



RESEARCH

NEED FOR RESEARCH

Some of the applications of information processing techniques described have existed for many years; others are barely beyond the research stage. The information management field, like most others, has made exceptional headway in some directions; additional work is needed in other areas that are currently the subject of a substantial research effort.

CLASSIFICATION AND INDEXING PROCEDURES

One series of SDC studies is examining the derivation of automatic and semiautomatic procedures for indexing, classifying, and abstracting documents. Initial results have shown that mathematically derived classification systems are useful in structuring large information repositories. Researchers are now analyzing new sets of documents to determine the reliability and consistency of factors previously derived and reported. Machine (i.e., automatic) classification of documents into derived categories is being compared with human classification.

As part of this work, SDC is studying the feasibility of representing the contents of files and libraries in "association maps" and "hierarchical maps." These representations, generated automatically on the basis of computer analysis of word associations in text, appear to have considerable promise as aids in information retrieval.

LINGUISTICS AND COMMUNICATION ANALYSIS

A second series of projects deals with linguistics and communication analysis. A much-used device in natural language discourse is the abbreviation of a long phrase in a given text sentence by a single word or short phrase in a later sentence of the text. This practice, while necessary for compact writing, produces difficulties in the machine processing of language. On the basis of a study of abbreviating phrases in a large corpus of scientific writing, tentative rules have been formulated for recognizing certain kinds of abbreviating phrases and locating the phrases in preceding sentences which they abbreviate. These rules are being coded for incorporation into mechanical paraphrasing and abstracting routines.

"FACT RETRIEVAL"

A third group of projects includes a study of "fact retrieval." An attempt is being made to ascertain the conditions at the interface between man and computer that will permit human requesters to communicate satisfactorily with a file of information elements. If adequate man-machine symbiosis can be achieved, it should be possible to supply the requester with the facts he needs, rather than with the document containing these facts. Some of the aspects being considered are the development of suitable request and storage languages, the formulation of complex information requests, and the potential of joint human-computer resolution of these problems.

NATURAL LANGUAGE PROCESSING

Extremely difficult to develop are computer programs that can process natural language information as if the meaning of the words were understood. Nevertheless, the information management field needs programs of this sort if it wishes to relieve the scientist, technician, businessman, or military commander of much of the burden of present-day information-handling procedures. One of SDC's research projects, known as Synthex, involves the development of computer programs that will enable a computer to accept a question in natural language, to analyze the syntax and the logical dependency relations of the question, to perform the functions necessary in searching the memory banks for the desired answer, and to output that answer in natural language.

To simulate human language behavior on a machine, Synthex must describe what a human does when he is presented with a question, finds or develops the answer from a relatively vast body of intellectual experiences, and states that answer. Something like this process has been systematized and translated into a complex program of machine instructions. The program has operated successfully on several types of literature and is now undergoing additional refinements to permit the drawing of inferences about information in the file. While the full application of such a capability is probably some years away, the present research is highlighting problems that must be solved in the meantime.

FUTURE OF INFORMATION MANAGEMENT

The next ten years are destined to stimulate even greater concern with information management. It is expected that by 1965 the expenditures on such systems will be more than 100 million dollars a year and will double every few years thereafter.

Advances in general- and special-purpose equipment will provide a stimulus for new information management applications. The cost per unit of information storage and processing equipment can be expected to decline and to bring powerful machine aids within the reach of small organizations.

New capabilities in information management techniques can also be anticipated. Systems will be developed that can provide information users with specific facts, rather than the body of published literature. These systems will help to select, compress, and organize material in order to conserve one of the nation's most valuable resources—its highly trained scientific and technical personnel.

The tremendous economic, scientific and technical progress since World War II, and the pressing need for accurate, high-speed, information management systems to support human decision-making, show no signs of tapering off. In the public interest, SDC will continue its efforts to contribute to the development of this important field.